Machine Learning 1

Lecture 12.4 - Kernel Methods
Gaussian Processes - With Exponential Kernels

Erik Bekkers

(Bishop 6.4.2)

Slide credits: Rianne van den Berg

Image credit: Kirillm | Getty Images

Drawing functions from GP's

- Specifying a kernel determines the characteristics over functions drawn from the GP

- For simplicity, let's take

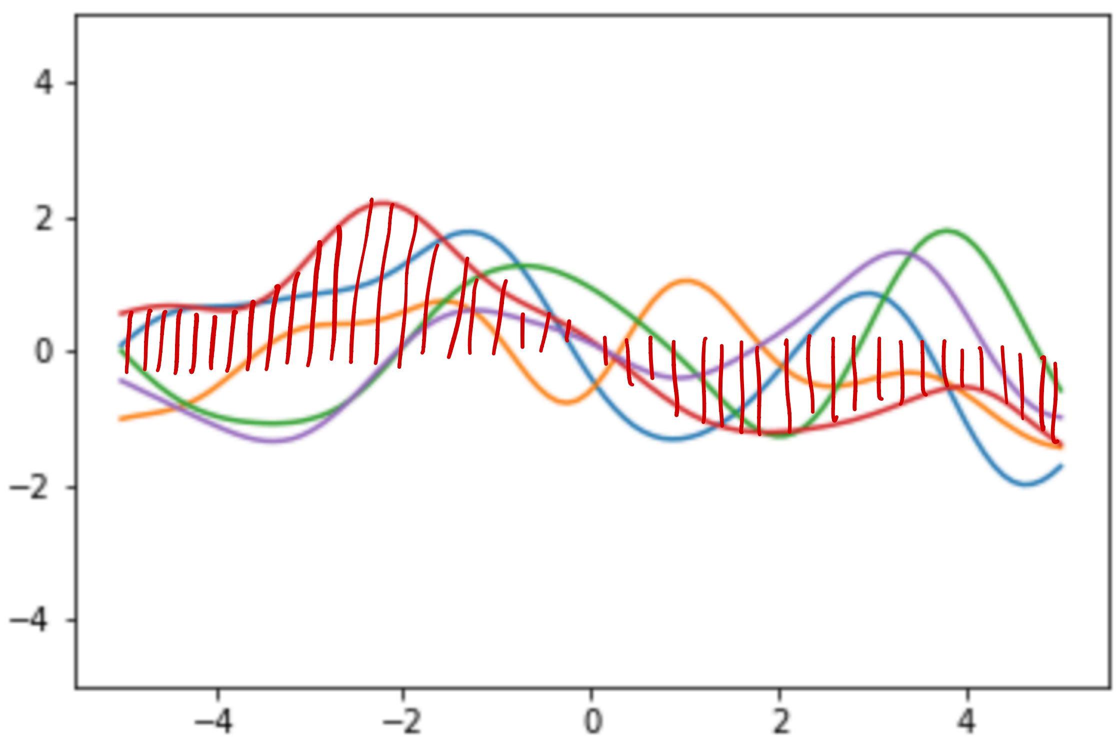
$$\mathbf{f} = \begin{bmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_N) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \dots & k(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) & \dots & k(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}\right)$$

- We consider this kernel

$$\boxed{k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 \underbrace{\exp\left(-\frac{1}{2\theta_1^2}||\mathbf{x}_n - \mathbf{x}_m||^2\right)}_{\text{exp}} + \underbrace{\theta_2 + \theta_3 \mathbf{x}_n^T \mathbf{x}_m}_{\text{"linear part"}}}$$

Drawing functions from GP's

- Sample fine grid of points $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \in [-5,5]$
- Compute $\mathbf{K}$
- Compute $\mathbf{K} = \mathbf{L}\mathbf{L}^T$ (Cholesky or eigen decomposition)
- Sample random vector of size N: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$
- Sample $\mathbf{f} = \begin{bmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_N) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{K}\right)$ by computing $\mathbf{f} = \mathbf{L}\mathbf{z}$



Using kernel:

$$k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 \exp\left(-\frac{1}{2\theta_1^2}||\mathbf{x}_n - \mathbf{x}_n||^2\right) + \theta_2 + \theta_3 \mathbf{x}_n^T \mathbf{x}_m$$

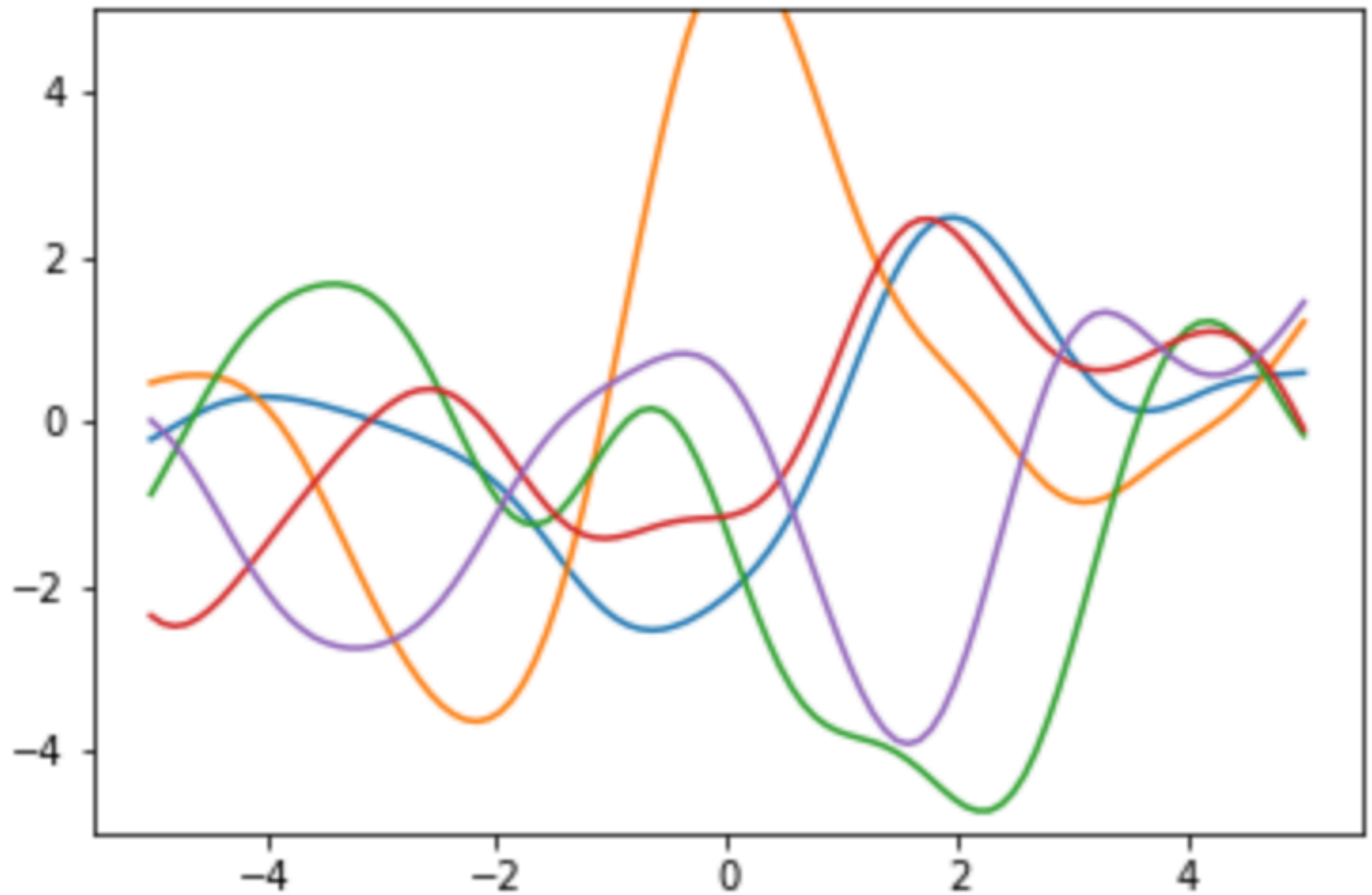
$\theta_0 = 1$
$\theta_1 = 1$
$\theta_2 = 0$
$\theta_3 = 0$

Varying the pre-factor θ_0

$$k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 \exp\left(-\frac{1}{2\theta_1^2}||\mathbf{x}_n - \mathbf{x}_n||^2\right) + \cancel{\theta_2 + \theta_3 \mathbf{x}_n^T \mathbf{x}_m}$$

Amplitude

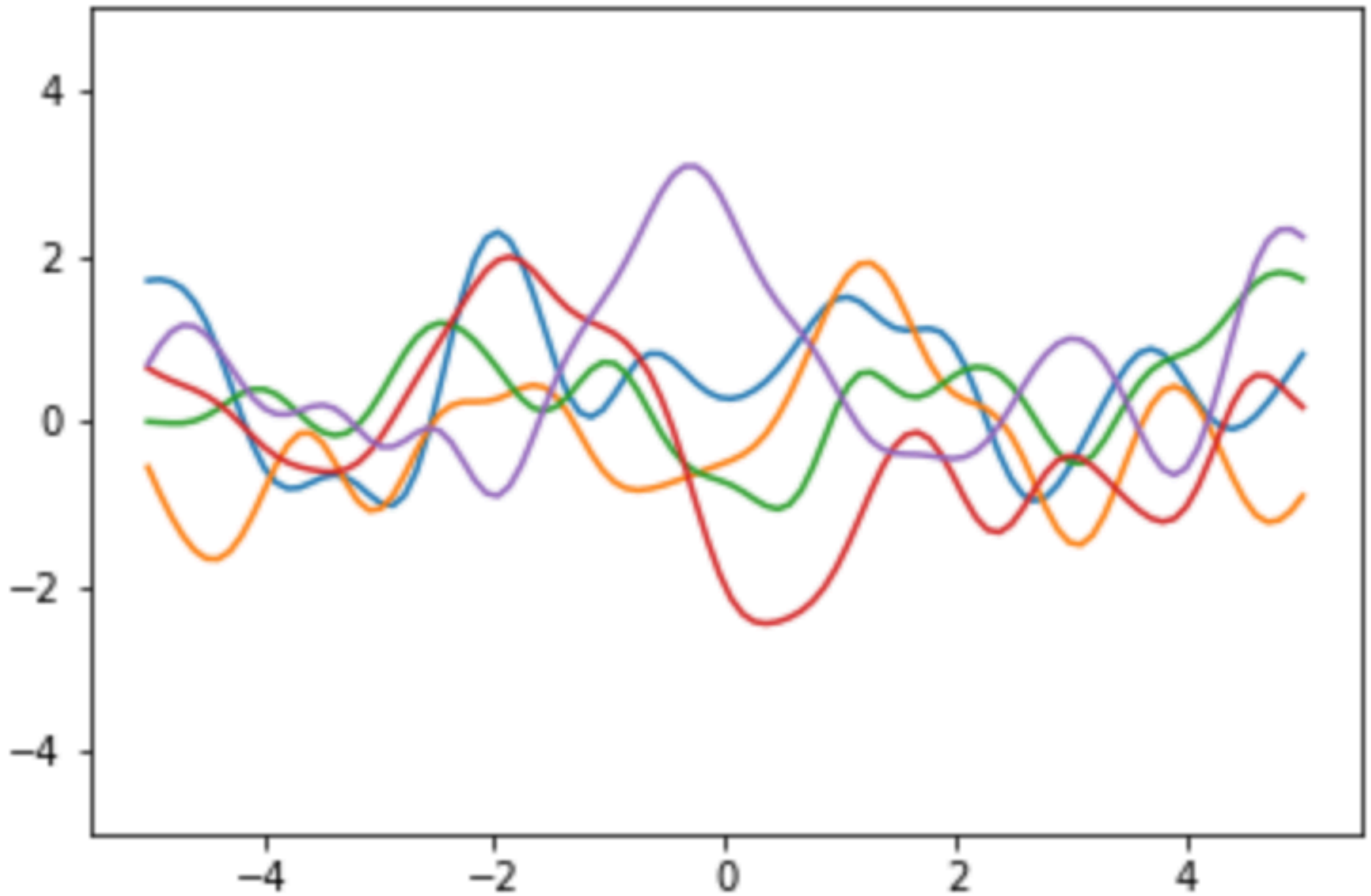
$\theta_0 = 4 \qquad \theta_1 = 1 \qquad \theta_2 = 0 \qquad \theta_3 = 0$



Varying the length scale θ_1

$$k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 \exp\left(-\frac{1}{2\theta_1^2}||\mathbf{x}_n - \mathbf{x}_m||^2\right) + \cancel{\theta_2 + \theta_3 \mathbf{x}_n^T \mathbf{x}_m}$$

$\theta_0 = 1 \quad \theta_1 = 0.5 \quad \cancel{\theta_2 = 0 \quad \theta_3 = 0}$



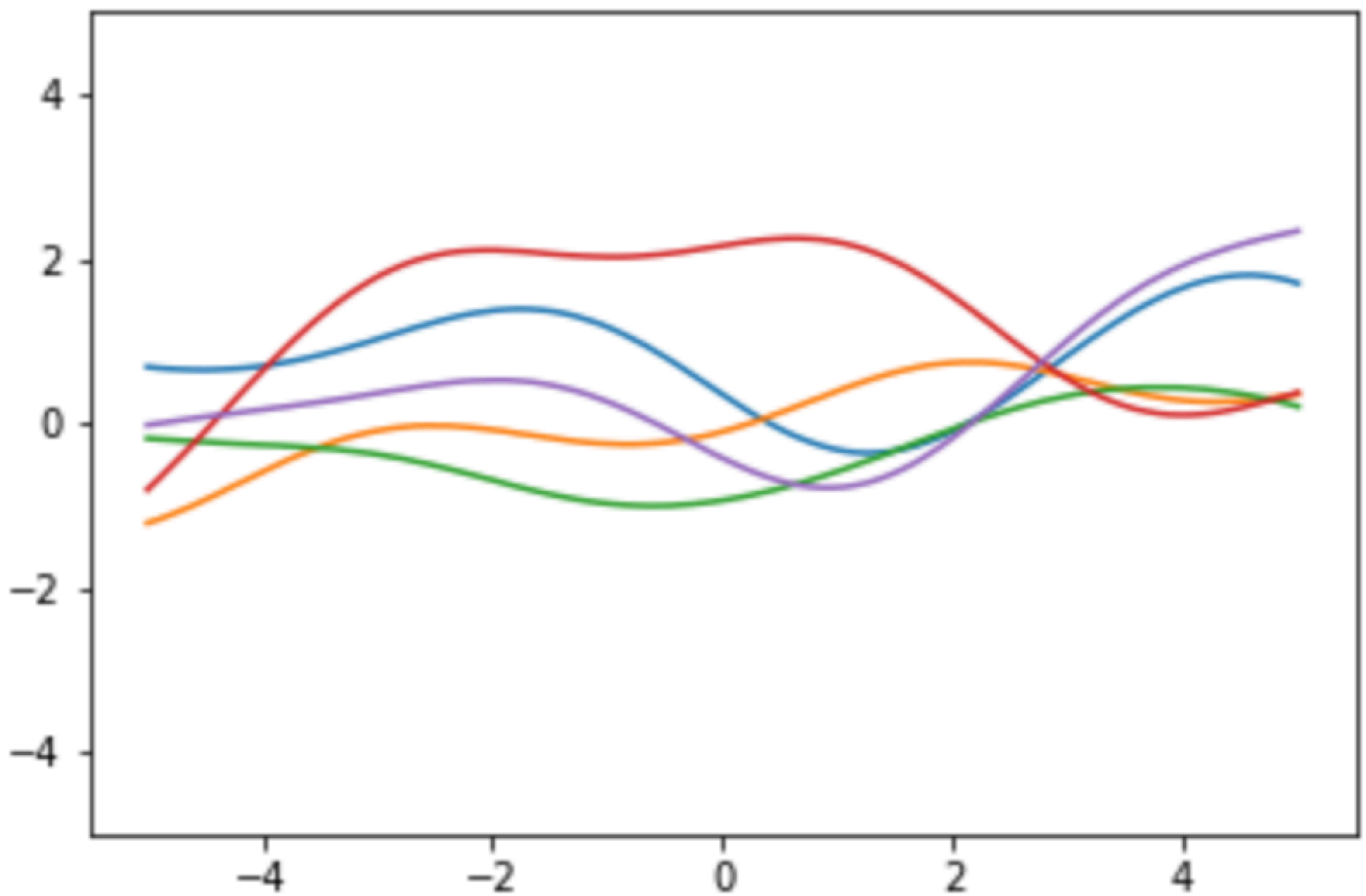
$\theta_1 \to 0$

$K \to I$

Varying the length scale θ_1

$$k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 \exp\left(-\frac{1}{2\theta_1^2}||\mathbf{x}_n - \mathbf{x}_m||^2\right) + \theta_2 + \theta_3 \mathbf{x}_n^T \mathbf{x}_m$$

$\theta_0 = 1 \qquad \theta_1 = 2 \qquad \theta_2 = 0 \qquad \theta_3 = 0$

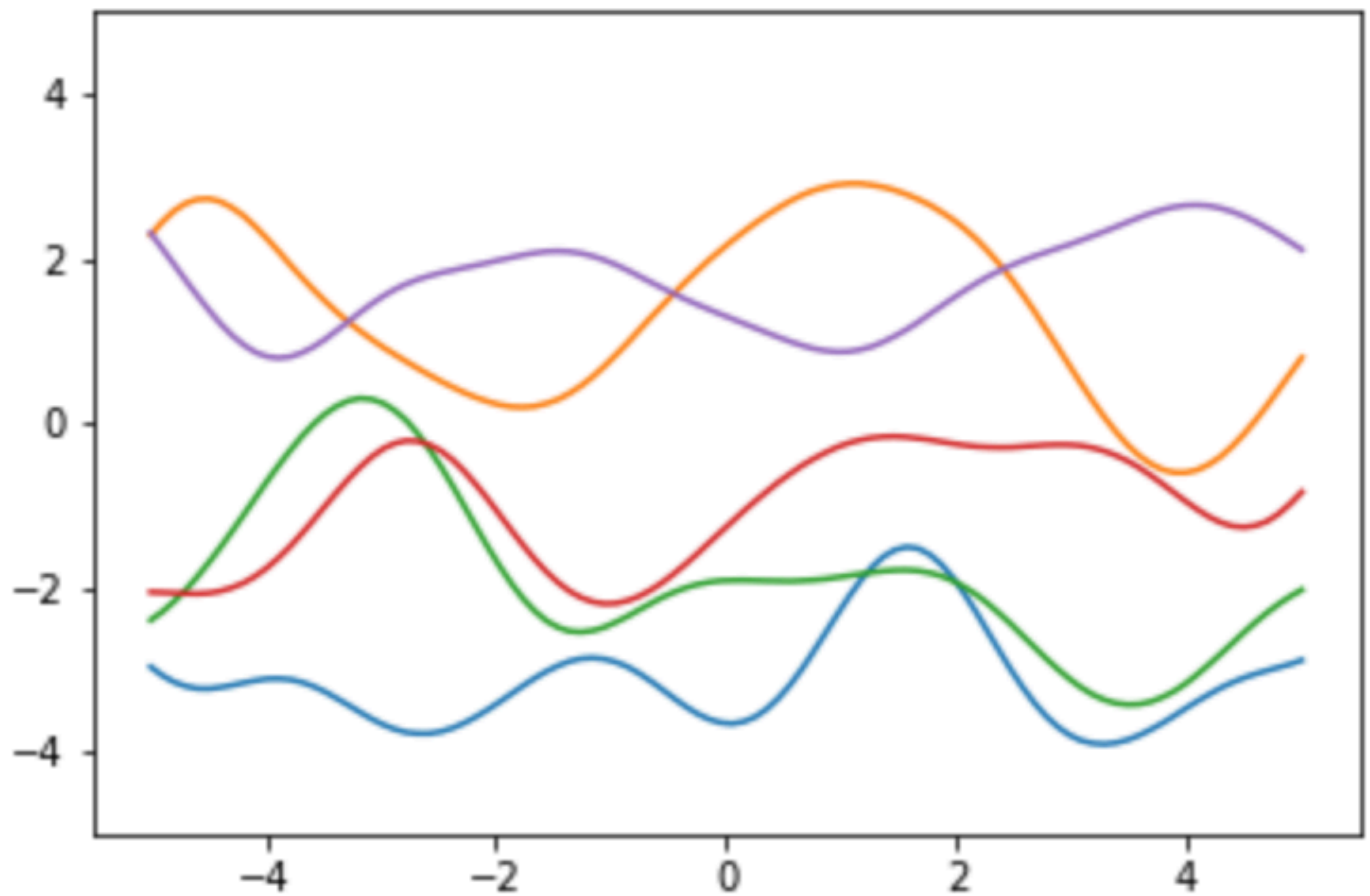


Varying the offset θ_2

correlation independent of position

$$k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 \exp\left(-\frac{1}{2\theta_1^2}||\mathbf{x}_n - \mathbf{x}_m||^2\right) + \theta_2 + \cancel{\theta_3 \mathbf{x}_n^T \mathbf{x}_m}$$

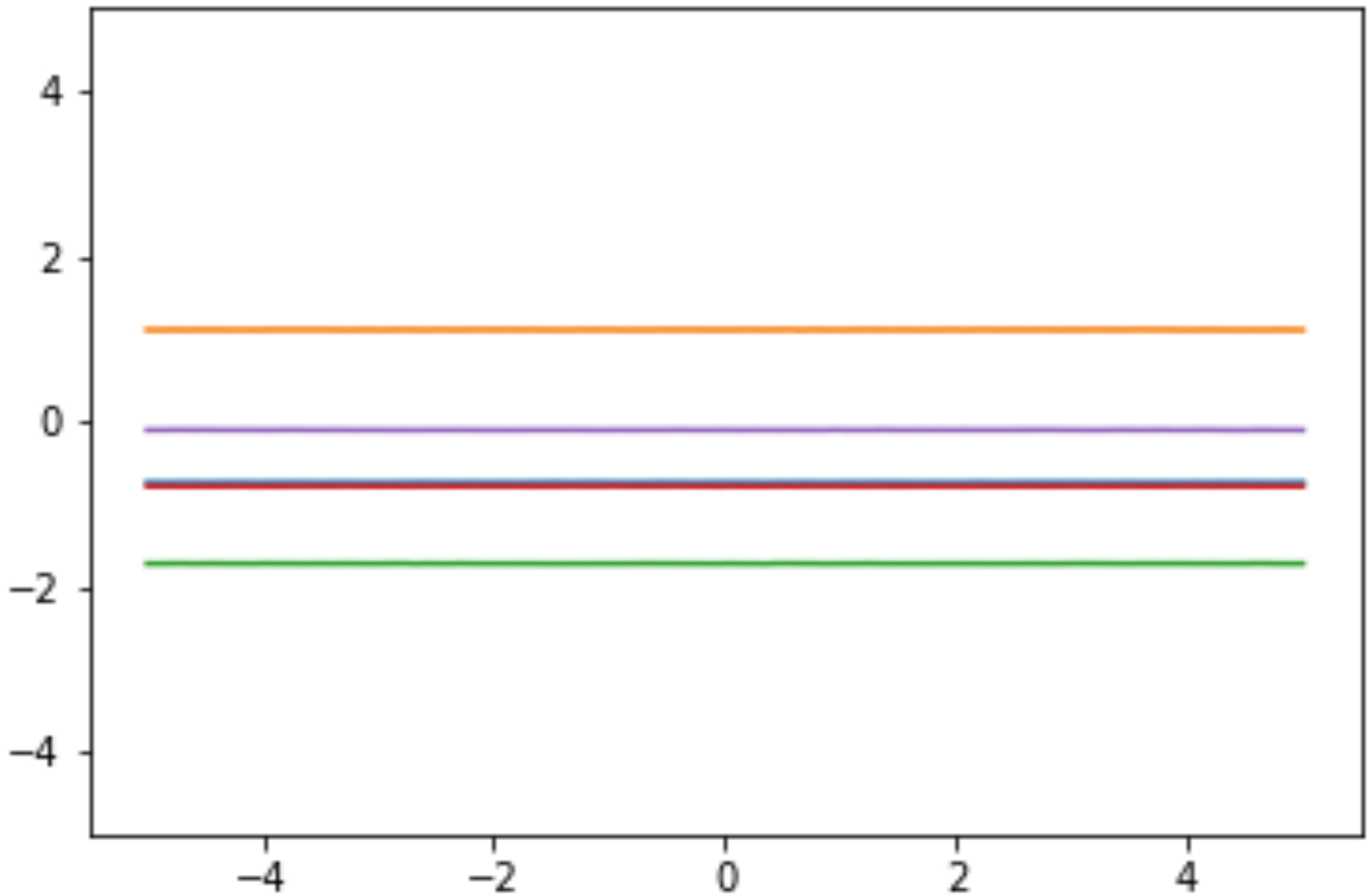
$$\theta_0 = 1 \qquad \theta_1 = 1 \qquad \theta_2 = 5 \qquad \theta_3 = 0$$



Varying the offset θ_2

$$k(\mathbf{x}_n, \mathbf{x}_m) = \cancel{\theta_0 \exp\left(-\frac{1}{2\theta_1^2}||\mathbf{x}_n - \mathbf{x}_m||^2\right)} + \theta_2 + \cancel{\theta_3 \mathbf{x}_n^T \mathbf{x}_m}$$

$$\cancel{\theta_0 = 0 \quad \theta_1 = 1} \quad \theta_2 = 5 \quad \cancel{\theta_3 = 0}$$

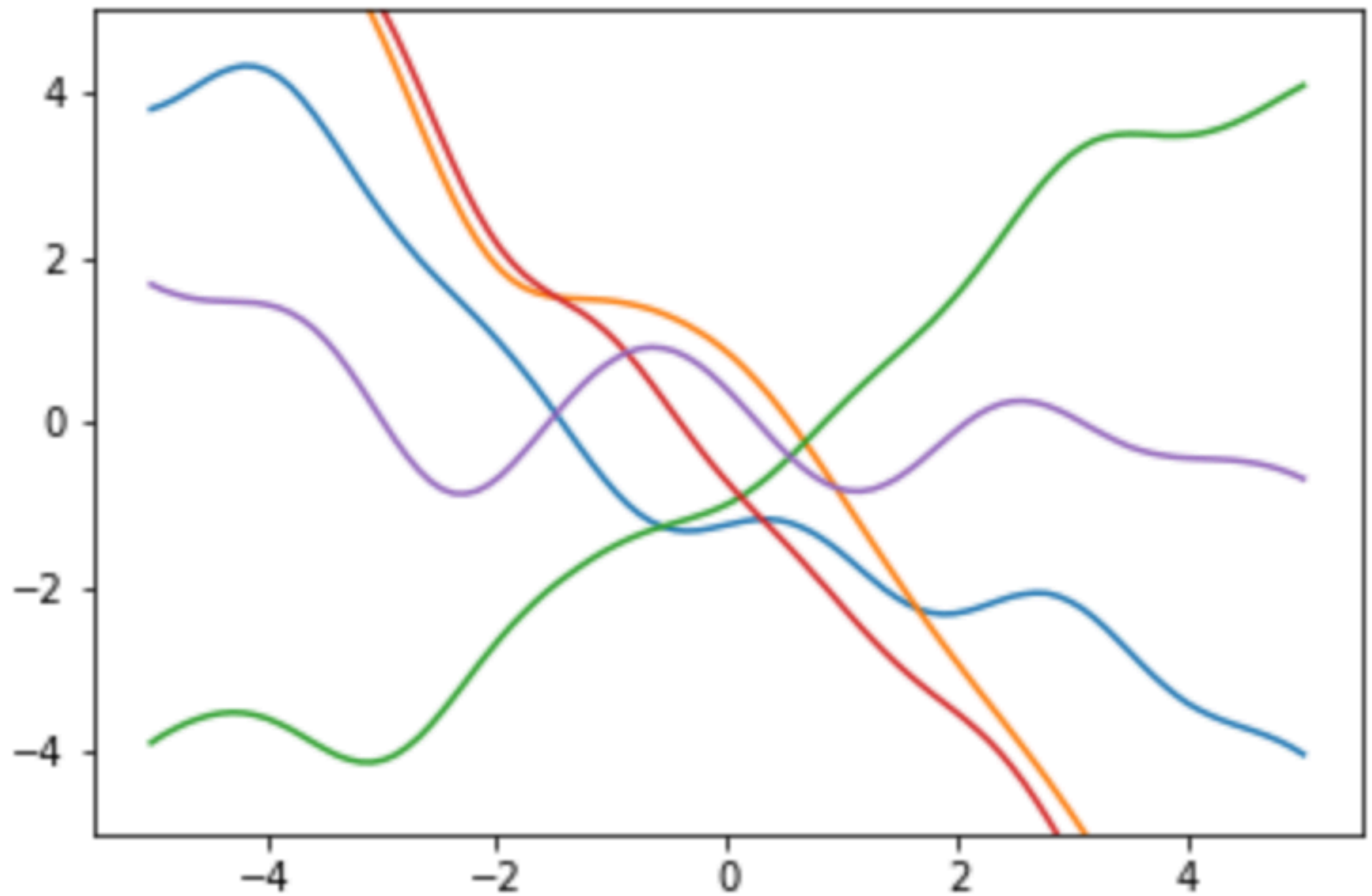


"perfect correlation"

Varying the linear term θ_3

$$k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 \exp\left(-\frac{1}{2\theta_1^2}||\mathbf{x}_n - \mathbf{x}_m||^2\right) + \theta_2 + \theta_3 \mathbf{x}_n^T \mathbf{x}_m$$

$$\theta_0 = 1 \qquad \theta_1 = 1 \qquad \cancel{\theta_2 = 0} \qquad \theta_3 = 1$$



Varying the linear term θ_3

$$k(\mathbf{x}_n, \mathbf{x}_m) = \cancel{\theta_0 \exp\left(-\frac{1}{2\theta_1^2}||\mathbf{x}_n - \mathbf{x}_m||^2\right)} + \cancel{\theta_2} + \theta_3 \mathbf{x}_n^T \mathbf{x}_m$$

$$\cancel{\theta_0 = 0} \quad \cancel{\theta_1 = 1} \quad \cancel{\theta_2 = 0} \quad \theta_3 = 0.2$$